Work Integrated Learning Programmes Division
M.Tech (AIML)
Machine Learning
S2-22_AIMLCZG565
Second Semester
Assignment 2 – PS-5

Breast Cancer Wisconsin (Diagnostic) Dataset [Weightage 10%]

Instructions for Assignment Evaluation

1. Please follow the naming convention as <Group no>_<Dataset name>.ipynb.
   Eg – for group 1 with a breast cancer wisconsin dataset your notebooks should be named as - Group1_BreastCancerWisconsin.ipynb.
2. Inside each jupyter notebook, you are required to mention your name, Group details and the Assignment dataset you will be working on.
3. Organize your code in separate sections for each task. Add comments to make the code readable.
4. Deep Learning Models are strictly not allowed. You are encouraged to learn classical Machine learning techniques and experience their behavior.
5. Notebooks without output shall not be considered for evaluation.
6. Delete unnecessary error messages and long outputs.
7. Prepare a jupyter notebook (recommended - Google Colab) to build, train and evaluate a Machine Learning model on the given dataset. Please read the instructions carefully.
8. Each group consists of up to 3 members. All members of the group will work on the same problem statement.
9. Each group should upload in CANVAS in respective locations under ASSIGNMENT Tab. Assignment submitted via means other than through CANVAS will not be graded.
10. Only two files should be uploaded in canvas without zipping them. One is ipynb file and other one html output of the ipynb file. No other files should be uploaded.

Problem Statement

Breast cancer is one of the most prevalent forms of cancer among women globally. Early and accurate diagnosis is crucial for effective treatment. Machine learning algorithms can play a significant role in assisting medical professionals in this regard. In this project, we aim to build and compare the performance of Random Forest and K-Nearest Neighbors (KNN) algorithms for predict the presence or absence of breast cancer malignancy using the Breast Cancer Wisconsin (Diagnostic) dataset.

Dataset: brcaw.csv (uploaded as separate file)

1. Import Libraries/Dataset

   a. Download the dataset.
   b. Import the required libraries.

2. Data Visualization and Exploration [1M]

   a. Print 2 rows for sanity check to identify all the features present in the dataset and if the target matches with them.
   b. Comment on class imbalance with appropriate visualization method.

c. Provide appropriate data visualizations to get an insight about the dataset.
d. Do the correlational analysis on the dataset. Provide a visualization for the same. Will this correlational analysis have effect on feature selection that you will perform in the next step? Justify your answer. **Answer without justification will not be awarded marks.**

3. Data Pre-processing and cleaning [2M]

a. Do the appropriate pre-processing of the data like identifying NULL or Missing Values if any, handling of outliers if present in the dataset, skewed data etc. Mention the pre-processing steps performed in the markdown cell. Explore few latest data balancing tasks and its effect on model evaluation parameters.
b. Apply appropriate feature engineering techniques for them. Apply the feature transformation techniques like Standardization, Normalization, etc. You are free to apply the appropriate transformations depending upon the structure and the complexity of your dataset. Provide proper justification. **Techniques used without justification will not be awarded marks**. Explore a few techniques for identifying feature importance for your feature engineering task.

4. Model Building [5M]

a. Split the dataset into training and test sets. **Answer without justification will not be awarded marks.** [1M]
    i. Train = 80 % Test = 20%
    ii. Also, try to split the dataset with different ratios of your choice.
b. Build a Random Forest classification model to predict the presence or absence of breast cancer. [2M]
    i. Tune hyperparameters (e.g., number of trees, maximum depth) using cross-validation. Justify your answer.
    ii. Evaluate the model's performance using appropriate metrics.
c. Build a KNN classification model to predict the presence or absence of breast cancer. [2M]
    i. Determine the optimal value of k through hyperparameter tuning and cross-validation. Justify your answer.
    ii. Evaluate the KNN model's performance using appropriate metrics.

5. Performance Evaluation [2M]

a. Compare the performance of the Random Forest and KNN models using appropriate evaluation metrics.
b. Provide insights into which model performs better and why. **Answer without justification will not be awarded marks.**